Modular Ontology Design Using Canonical Building Blocks in the Biochemistry Domain

Christopher J. THOMAS[1], Amit P. SHETH[1] and William S. YORK[2]

[1] LSDIS Lab, Department of Computer Science, University of Georgia
[2] Complex Carbohydrate Research Center (CCRC), University of Georgia
Athens, Georgia, USA

{cthomas, amit}@cs.uga.edu, will@ccrc.uga.edu

Abstract

The field of BioInformatics has become a major venue for the development and application of computational ontologies. Ranging from controlled vocabularies to annotation of experimental data to reasoning tasks, BioOntologies are advancing to form a comprehensive knowledge foundation in this field. With the Glycomics Ontology (GlycO), we are aiming at providing both a sufficiently large knowledge base and a schema that allows classification of and reasoning about the concepts we expect to encounter in the glycoproteomics field. The schema exploits the expressiveness of OWL-DL to place restrictions on relationships, thus making it suitable to be used as a means to classify new instance data. On the instance level, the knowledge is modularized to address granularity issues regularly found in ontology design. Larger structures are semantically composed from smaller canonical building blocks. The information needed to populate the knowledge base is automatically extracted from several partially overlapping sources. In order to avoid multiple entries, transformation and disambiguation techniques are applied. An intelligent search is then used to identify the individual building blocks that model the larger chemical structures. To ensure ontological soundness, GlycO has been annotated with OntoClean properties and evaluated with respect to those. In order to facilitate its use in conjunction with other biomedical Ontologies, GlycO has been checked for NCBO compliance and has been submitted to the OBO website

Introduction

The field of BioInformatics has seen a dramatic increase of available ontologies for many of the life sciences domains. The Ontologies in the OBO project [17], especially the Gene Ontology (GO) [6] with its comprehensive schema and thousands of instances, take leading roles. As a broad lexicon or dictionary, GO serves one of the major purposes of ontologies: facilitating agreement. However, it is not designed for extensive computational use, so the amount of machine-accessible knowledge is limited. Only two types of relationships between the different entities in the ontology are formalized, *is_a* and *part_of*. Other relationships can only be simulated by reification of new terms that are then used in the *is_a* and *part_of* hierarchies [22].

An ontology that provides rich, machine accessible relationships must be formalized. Knowledge modeling languages such as KIF [7], RDF [13] or the W3C-recommended Ontology Web Language OWL [11] allow such formalizations with different expressiveness. OWL in its three flavors Lite, DL and Full promises to be a good compromise between expressiveness and computational complexity on the one hand and versatility and simplicity on the other.

In the context of the “BioInformatics for Glycan Expression” core of the NCRR Integrated Resource for Biomedical Glycomics project, a suite of web-accessible ontologies has been developed for the glycoproteomics domain. The goal of this suite is to have a basis for description, annotation and reasoning, such that every step from experimental setup over experimental conduct and analysis to acquisition of hypotheses and theories can be formalized. This paper focuses on issues related to representation, expressiveness, granularity and instance population in the development of the Glycan Structure Ontology GlycO.

Glycans are complex carbohydrate structures, which play key roles in the development and maintenance of living cells. Glycans are built from simpler monosaccharide residues (such as mannose and glucose), which constitute the nodes of tree structures with edges that are comprised of chemical bonds between the residues. The synthesis of these glycans in organisms is an intricate process that can be modeled as a collection of biosynthetic pathways. At each step in such a pathway, an enzyme-catalyzed reaction ‘adds’ a new residue as a leaf to an existing structure or ‘moves’ a whole sub-tree to a different parent. It is well established that alongside genes and proteins, glycans play a major role in cell functions.

The aim of glycoproteomics is to understand cellular processes that are mediated by the interaction of proteins, the genes that encode them, and the glycans that are attached to their surfaces. Our goal in developing GlycO has been to assess the extent to which knowledge in this domain can be logically formalized to facilitate the discovery and specification of relationships between the glycan structures, their metabolism, and their functions. Among the challenges faced were those of a limited expressiveness of the chosen OWL-DL standard, and mereological issues of granularity.

The main contributions of this work include:

- Creating a more meaningful domain model by
    - Building a schema that captures the richness of the domain using expressive language, esp. restrictions
    - Supporting modeling of molecular structures that are important for domain scientists
    - Rigorously modeling with canonical instances used as building blocks
- Populating the ontology by extracting and disambiguating instance information from multiple heterogeneous sources
- Allowing for more meaningful queries by formalizing knowledge that is usually inferred in database models
- Addressing granularity issues

Following this introduction, section 2 will describe the conceptualization and formalization of the glycoproteomics domain in GlycO. section 3 will detail the sources and algorithms used for the automatic population, while section 4 will evaluate GlycO and discuss the impact it can have on biochemical applications. Section 5 finally concludes the paper.

2. Ontology Design

2.1 General Considerations

The rules of syntax alone cannot determine the meaning of the statements expressed by the words in that syntax. A fundamental aspect of ontology development is the capture of semantics in a formal syntax, i.e., the unambiguous formalization of statements or states of affairs. Representation of meaning using first order logic is limited to stating that an object has certain properties and relationships with other objects. Even generalizing these properties to sets or classes of objects bears problems [22]. It is necessary to find a balance between the unambiguous representation of objects including their relationships and any attempt to capture the infinitude of relationships present in the world.

We therefore are limited to modeling very specific problems that require a finite amount of representation. The critical objects and their relationships must be identified and then formalized so that machines can infer new or implicit knowledge from the given information. Despite the identified fact that syntax incompletely determines semantics, in cases of restricted domains the actual words and their order in a statement can correspond quite directly to the meaning of it. Hence, if we know the rules that govern the syntax as well as the context, the words and their syntactical structure often suffice to determine their meaning.

Collections of biological entities, such as genes, proteins and carbohydrates, are assumed to have a syntactic structure, much like natural language. For example, we assume that the structure of the genome directly or indirectly encodes the structure of the entire organism. By knowing the syntactic and semantic rules that govern gene structure, we can assign meanings to DNA strings and substrings, *i.e.*, identify genes and the protein sequences they encode. Of course, this is not always a trivial task, but provided the genes themselves (and not their environmental context) constitute the information basis, we can gain a large amount of knowledge by studying gene syntax and semantics. We make a similar simplifying assumption for glycans, which clearly influence cellular properties. Ideally, we can capture the correspondence between a glycoprotein's biological properties and the presence of specific glycan structures at specific locations on the protein's surface.

Developing a highly expressive formal ontology for a comparatively narrow field of research requires the constant interaction between domain experts and knowledge engineers. The modeling of knowledge calls for a profound understanding of a domain. The domain expert must fully participate in ontology development and understand the formalisms used for specifying the conceptualization of the domain. Conversely, the knowledge engineer must analyze the ontology to avoid ontological fallacies in modeling. The Ontoclean methodology [9] explains how concepts should be classified on a meta-level according to distinctions like rigid versus non-rigid concepts, entities versus roles, etc. The knowledge engineer must have enough domain knowledge to apply these distinctions to the ontology.

Although GlycO is focused on the glycoproteomics domain, it is critical that it is sufficiently comprehensive to invoke important concepts in the related disciplines of proteomics and genomics. By providing links to other ontologies that describe the fields closely related to glycoproteomics, it allows for scientific discovery of complex or unknown relationships across research fields. Because it is assumed that the ontology will be used for such discovery, it needed to be strongly restricted to clearly distinguish the asserted concepts by semantically modeling the subtle differences in glycan structure that modulate their biological functions. Only then a correct identification of discovered concepts and relationships can be achieved. GlycO is meant to be more than a controlled vocabulary; its intention is to be used for reasoning in scientific analysis and discovery.

2.2 Schema Design

Initially, the glycoproteomics domain was broadly analyzed, terms were collected, and the way these terms are used by scientists was examined. It turns out that the informal usage of the *is_a* relationship, as in "a glycan is a carbohydrate", implies a hierarchy of concepts with multiple inheritances. We wanted to keep the "colloquial" use of the biochemistry terminology consistent with the ontology, while also adding more distinguishing descriptions in the form of named relationships and their restrictions. There are many ways of classifying monosaccharide residues, which are the building blocks of glycans. For example, it is possible (and equally valid) to classify them according to the number of carbon atoms in the monosaccharide or as a structural variant. That is, a β-D-Glc*p* residue can be identified amongst other criteria as both a hexosyl residue (with 6 carbons) and an aldosyl residue (embodying the aldo- structural variant). We account for all of these properties by allowing a particular monosaccharide residue to inherit from several super classes. Whether this directed acyclic graph is explicitly asserted or subsequently inferred is secondary. For example, the absolute configuration D and subsumption by the superclass *residue* are necessary and sufficient properties of the class D-*residue*. A reasoner will automatically subsume any *residue* class that has the absolute configuration D under the class D-*residue*. A hierarchy with multiple inheritance will almost always automatically arise when a more sophisticated logical description of classes is used alongside restricting conditions. For this reason, criticism of multiple inheritance, as in [23] seems impractical to us.

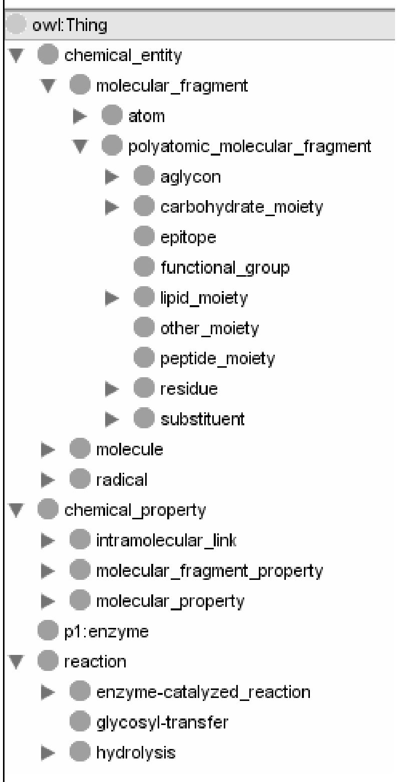
The first level of abstraction contains the three classes "*Chemical Entity*", "*Chemical Property*" and "*Reaction*". This is an appropriate starting point in that upper level ontologies such as SUMO distinguish between "Object", "Attribute" and "Process". The Gene Ontology uses *cellular_component*, *biological_process* and *molecular_function* on the first level of abstraction. The analog to *molecular_function* is in our case defined in the functional ontology *EnzyO [4]*, which describes enzymes and their functions. This compliance with standard classifications facilitates the integration of GlycO with other ontologies. From there, a finely grained class hierarchy is defined (see Figure 1 for a selection of the first 4 levels of the GlycO hierarchy).

The relationship hierarchy in GlycO is built with respect to emerging standards in the biomedical domain. The OBO relationship ontology [17][22] is used as a starting point and more refined named relationships are added. See Figure 2 for a part of the GlycO relationship hierarchy.

With 14 levels, GlycO has a deeper hierarchy than many other domain ontologies. This finely grained class design is essential for the purposes of evaluating experimental results using the knowledge stored in the ontology. Small differences in the glycan structure might affect the kind of interactions an individual glycan or members of a class of glycans have with other objects in the ontology.

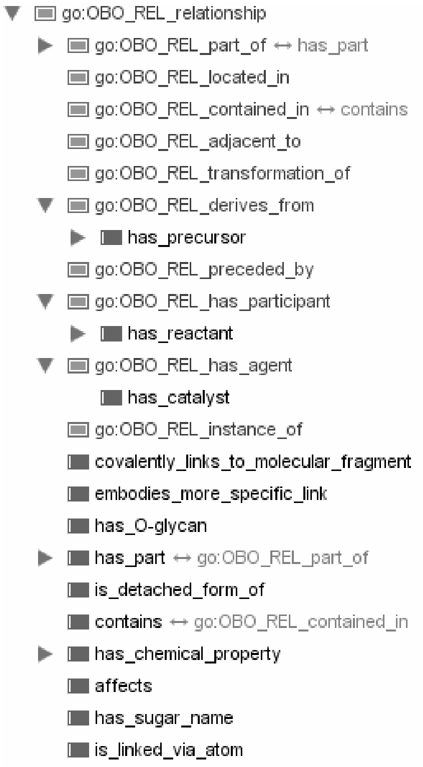
The hierarchy of concepts is one aspect of semantics captured in an ontology, but the addition of other relationships is required to realize an expressive model. A concept by itself might be useful for a human observer, but only by understanding it within a context of other concepts. Scientists infer related concepts according to their background knowledge. For machines, this background knowledge needs to be stated explicitly. The authors of [23] raised the issue that the biomedical ontology MGED contained too many named relationships that impede the computational use of the ontology. We disagree with this assessment of ontology design. A large number of named relationship increases the semantic value of an ontology [21], if these relationships are well defined. We address the dilemma of generality versus computational complexity by making use of a relationship hierarchy, modeling the relationships from more general down to more specific. Upper level relationships are e.g. *has_part* or *affects* and their inverses. Inheriting lower level relationships restrict domains and ranges of the upper level relationships. For example, *has_carbohydrate_residue* is essentially a *has_part* relationship, but its domain is restricted to *glycan* and its range is restricted to *carbohydrate_residue*. If the ontology is to be merged or aligned, an alignment algorithm will be able to map this relationship to a more general relationship in a different ontology that does not explicitly formalize the specific *has_carbohydrate_residue* relationship.

As the name indicates, a class hierarchy provides a means of classification. Together with relationships and restrictions it specifies what can possibly exist within the realm that is described. Classes themselves exist only in a very abstract sense. The instances in the ontology are meant to provide a representation of the things that actually exist in the domain of interest.



```
owl:Thing
▼ chemical_entity
    ▼ molecular_fragment
        ▶ atom
        ▼ polyatomic_molecular_fragment
            ▶ aglycon
            ▶ carbohydrate_moiety
              epitope
              functional_group
            ▶ lipid_moiety
              other_moiety
              peptide_moiety
            ▶ residue
            ▶ substituent
    ▶ molecule
    ▶ radical
▼ chemical_property
    ▶ intramolecular_link
    ▶ molecular_fragment_property
    ▶ molecular_property
  p1:enzyme
▼ reaction
    ▶ enzyme-catalyzed_reaction
      glycosyl-transfer
    ▶ hydrolysis
```

Figure 1: Selection of the first 3 levels in the GlycO hierarchy



```
▼ go:OBO_REL_relationship
    ▶ go:OBO_REL_part_of ↔ has_part
      go:OBO_REL_located_in
      go:OBO_REL_contained_in ↔ contains
      go:OBO_REL_adjacent_to
      go:OBO_REL_transformation_of
    ▼ go:OBO_REL_derives_from
        ▶ has_precursor
      go:OBO_REL_preceded_by
    ▼ go:OBO_REL_has_participant
        ▶ has_reactant
    ▼ go:OBO_REL_has_agent
          has_catalyst
      go:OBO_REL_instance_of
      covalently_links_to_molecular_fragment
      embodies_more_specific_link
      has_O-glycan
    ▶ has_part ↔ go:OBO_REL_part_of
      is_detached_form_of
      contains ↔ go:OBO_REL_contained_in
    ▶ has_chemical_property
      affects
      has_sugar_name
      is_linked_via_atom
```

Figure 2: Part of the relationship hierarchy

2.3 Canonical Instances

The problem of deciding where to make the cut between classes and instances and what to consider as an instance is well known in ontology design [16]. Even though *OntoClean* [9] describes some fallacies that can occur when making wrong choices for classes vs. instances, it is usually seen as an arbitrary, domain- or task-dependent choice. There is no rigorous formal methodology behind these choices.

Noy and McGuinness [16] give a good example for the wine ontology in which the designer has to decide whether the type of wine or the single bottle are of particular interest to the users of the ontology.

By analogy, an ontology in the glycan domain could describe individual glycan molecules. With 10^{15} (or more) chemically identical glycan molecules in a purified laboratory sample, this would be a tedious and useless endeavor. It makes much more sense to describe archetypal glycan molecules. Within the context of GlycO, it is not very useful to have a simple, mostly textual description of the glycan structure, as in most carbohydrate databases. To describe the complex structural features of glycans, each glycan is composed of several building block instances that model the monosaccharide residues. Each residue instance is richly described by the sub-tree it terminates and by additional properties that define how it is chemically linked to the next residue in the glycan. We chose this level of granularity for our description because these individual features can be associated with the physiological properties of the glycan and the cellular machinery involved in its biosynthesis, catabolism, recognition, etc.

For the current version, which focuses on the N-glycans subclass, this is accomplished by defining a tree structure of canonical residue entities that subsumes most N-glycans. That is, almost all of the known N-glycan structures can be completely specified by choosing a subset of the nodes of this tree. This subset forms a connected subtree that includes the root residue. This tree (known as GlycoTree) has been previously described [25], and we have formalized that structure as a collection of interconnected, canonical residue instances in GlycO. See Figure 3 for an image of GlycoTree.

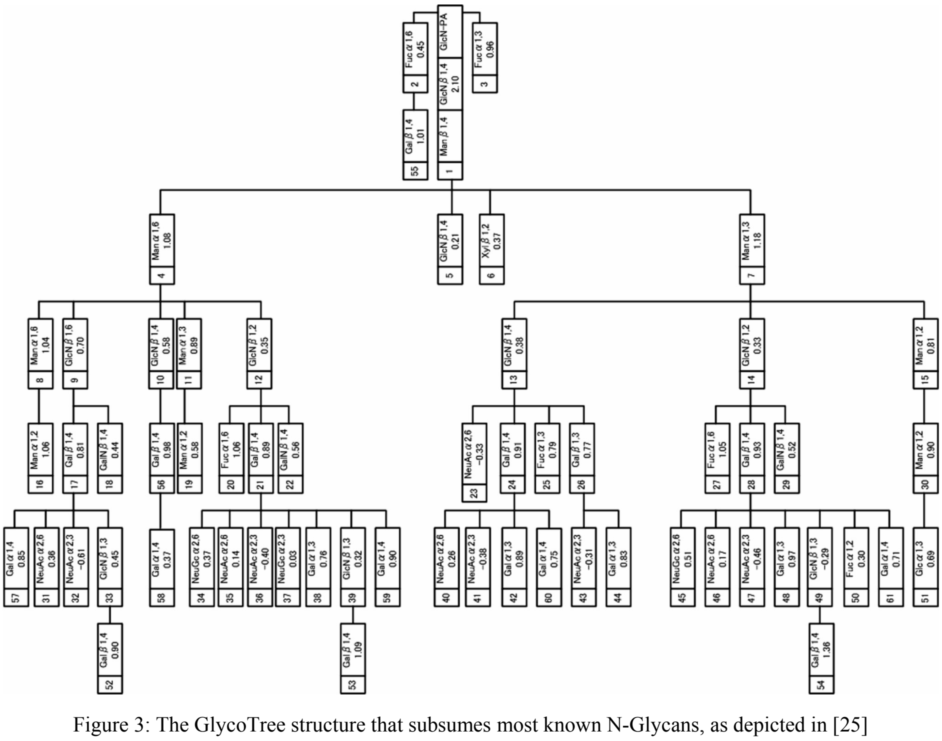


Figure 3: The GlycoTree structure that subsumes most known N-Glycans, as depicted in [25]

In spite of its practicality, the use of canonical residues to describe glycan structures evoked some ontological problems. If a glycan instance is chosen as a representative for all real glycans that have this structure, can we also let a residue instance that appears in many glycan instances, be at that same level of abstraction? The key question here was in our case to which extent an instance is determined by its context. In particular, the issue was whether it was ontologically justifiable to have each residue instance determined only by its chemical structure and the residue to which it is linked in the glycan. From a purely structural point of view this was justified with the GlycoTree structure elaborated by Takahashi and Kato [25]. Practically, it is justified by the reduction in the number of residue instances that results when different glycans can "reuse" the same residue in the same position. We believe that we can also semantically justify this decision because it reflects the way glycans are synthesized along their metabolic pathways, where enzyme-catalyzed reactions 'add' new residues as leaves to the existing glycan tree structures or 'move' the entire glycan to a protein. A specific type of residue is added in a reaction catalyzed by a specific enzyme at a specific position in the precursor glycan. We know that, for example, a mannose residue in position 1 is functionally different from a mannose residue in position 4. What remains to be demonstrated is whether residues in the same position in different glycans can be mapped to a particular function or participation in a metabolic pathway. This assumption is naturally underlying the current implementation. The chosen design can help determining whether this assumption is valid or not, because it is easily falsifiable on a case-by-case basis. We can easily establish sets of glycans that contain the same canonical residue instance and query whether the members of the set have common biological functions or are part of the same metabolic pathway.

Another issue of granularity is deciding which granular partitions of the world are represented [1]. Even in the molecular context of GlycO, different levels of granularity arise, especially when it comes to the representation of chemical linkage. Conceptually, larger molecular fragments are linked together, for example in glycans that attach to proteins. However, the actual link is naturally between two atoms. Intermediate links can also be asserted, such as the link between the glycan root residue and the amino acid in the protein that it attaches to. This issue was resolved by allowing chemical links to embody all these links recursively. The link is promoted from a simple relationship to a first class object that is defined by the two objects it links and by a more refined link.

Furthermore, atoms are parts of molecular fragments, which in turn are parts of molecules. This is an example of a partition into bona-fide versus fiat objects [1]. Molecules exist as wholes independently of other objects Molecular fragments describe functional partitions, even though they actually exist as such for extremely short amounts of time during chemical reactions, and should thus rather be seen as fiat objects.

3. Populating the Ontology

3.1 General Considerations

Creating ontologies is usually costly. In addition to a schema design, the actual domain knowledge in form of instances needs to be gathered, conceptualized and formalized. CYC [14] and GO are examples of ontologies that require high maintenance, due to the need for manual curation. This is not an issue in ontologies that only describe a schema to be used for database integration or as vocabularies. But since instance descriptions in GlycO are very different from those found in databases, ways to automate this process needed to be found. The objective in the development of GlycO was to have an expressive and restrictive schema that allows automatic and hence less expensive maintenance, given that semi-structured and reliable information is available for its population.

3.2 Populating GlycO from trusted sources

With CarbBank [3], KEGG [12] and SweetDB [15], several databases exist that contain trusted and up-to-date information about glycan structures. Even though CarbBank was discontinued, its content is of high quality and it is still used as a reference in other databases.

The GlycO schema specifies more complex relationships than these databases. A large number of properties not specified in their schema can be computationally inferred from the information given in the databases and are then explicitly added to the glycan description in the ontology. Hence we use these sources to populate the ontology with carbohydrate instances, alongside other sources for the population of gene and protein information. We assume that while each of the databases can contain incorrect entries, it is less likely that all three have the same incorrect entry. For this reason we extracted information from all these databases and compared this information during the population. To gather the data, the Semagix Freedom toolkit[20] was used that facilitates extraction of information from semi-structured websites and converts it to a structured representation that can be exported as XML or RDF or accessed via an API.

3.3 An Intelligent Population Algorithm

A structured representation of data does not necessarily guarantee its usefulness. Since the information was extracted from different sources, it has to be disambiguated to avoid having differently named copies of the same structure. As mentioned above, a simple textual description of structures is not suitable for our purposes and would only give an RDF encoding of already existing databases. In order to disambiguate the potential instances, the textual description of the structure was converted into the internal GlycoTree representation. This was performed using a multi-step process in which ambiguity is progressively removed as more meaningful representations are generated.

Conventionally, glycans are represented in the so-called IUPAC format, which is a two-dimensional textual representation that visually reflects the inherent tree structure and is easily comprehended by the human eye. Unfortunately, this representation is not unique. A web service is provided[1] that converts this representation into the structurally unambiguous LInear Notation for Unique description of Carbohydrate Sequences (LINUCS) [2]. Since this conversion is purely based on structure, it does not disambiguate different naming conventions for the substructures of the complex carbohydrate, the monosaccharide residues. For this purpose, another conversion is used that transforms the LINUCS representation into the XML-based GLYcan Data Exchange (GLYDE) format [19], which semantically disambiguates the different naming conventions of monosaccharide residues. XML has an inherent tree structure and GLYDE uses this fact. A child monosaccharide residue in a glycan is simply represented as a child node in the XML representation. This makes it relatively easy to perform tree operations on this representation. (See Figure 4 for the population workflow)

In the GlycoTree model each monosaccharide residue is defined by its type, its linkage and its position in the GlycoTree. Because of its canonical representation, the root node

[1] http://www.glycosciences.de/tools/linucs/

of a glycan can potentially be the root node of any sub tree of the GlycoTree. The population algorithm identifies and assigns the sub tree that corresponds to a particular glycan that is to be instantiated in the ontology. This is done by looking for sub tree isomorphisms. Several efficient sub tree isomorphism algorithms are available [18]. In our case, because of comparable small glycan structures, a depth-first search was sufficient. Additionally, the glycan constitutes a complete sub tree isomorphism; i.e. there cannot be a node in the glycan representation that is not part of the larger tree, nor can there be merely a homomorphism such that edges in the GlycoTree would need to be contracted to accommodate the glycan structure. If no isomorphism can be found, new GlycoTree nodes are generated automatically to complete the ontology. Here as well a report is generated so the domain expert can verify the correctness. New tree nodes can be inappropriately generated as a result of an incorrect structural description or classification of the glycan in the database. We identified several incorrect glycan descriptions by checking all new nodes that were generated during the population process. As only a few new nodes were generated, this is much easier than checking the entire set of glycan instances for errors.

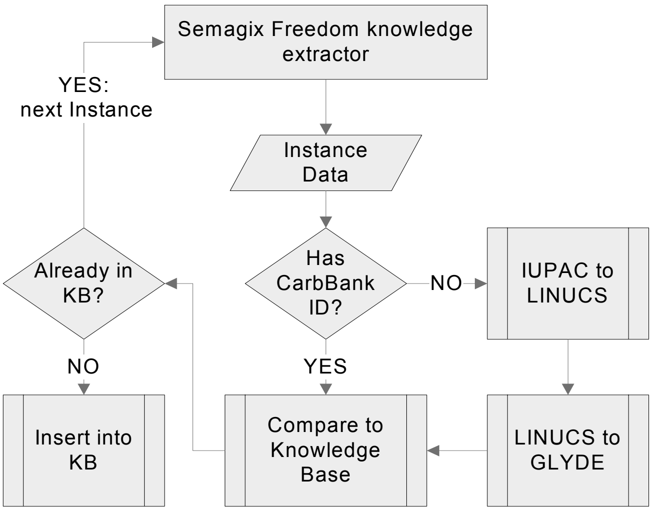


Figure 4: GlycO population workflow

The population algorithm will also be used to automatically build minimal trees for other glycan subclasses, such as O-glycans and glycolipids, which have not been classified entirely in such a tree structure. In [10] such tree structures are built, but only cover 61.2% of the known carbohydrate structures.

The set of GlycoTree nodes that represent a particular glycan can be easily compared to another set of nodes that represents a different glycan instance in the ontology. Two glycans are the same if and only if their tree node sets are identical. This method of disambiguation proved to be the more robust than other criteria, such as a common identifier, which is unreliable because every database uses proprietary accession numbers. Although all of the databases that were used as trusted sources make reference to CarbBank identifiers, CarbBank is no longer actively curated and these databases contain glycans that do not have a CarbBank ID.

4. Evaluation

It is difficult to measure the quality of an ontology. Guarino [8] proposed an evaluation based on precision and recall with respect to a reference conceptualization. This of course requires a formal conceptualization that applies to the same domain. With respect to the OntoClean ontology, for example, such a formal evaluation can show whether certain meta-properties of concepts are correctly assigned in the ontology. We rigorously modeled the GlycO ontology according to this meta-methodology. Another dimension for evaluation are structural metrics that assign numerical values to criteria such as depth, breadth, fan-outness, etc. [5][26]. These metrics are useful especially in large ontologies to get an idea of their structural character. Of course, none of these metrics can really tell us how useful an ontology will be and how well it models its domain. Table 1 shows the results of comparing GlycO to other biomedical ontologies using these metrics. Instance information is not taken into consideration. GlycO shows the highest connectivity, indicating a rich set of well defined and logically restricted relationships. The average number of sub terms gives an indication of the fan-out, but also the depth of GlycO. In a comparable fan-out measure, when siblings are counted, the number of siblings ranges between 1 and 15 with an average of 6.

Ontology	*No. of Terms*	*Avg. sub- terms*	*Connectivity*
GlycO	**324**	**2.5**	**1.7**
ProPreO	**244**	**3.2**	**1.1**
MGED	228	5.1	0.33
Biological Imaging methods	260	5.2	1.0
Protein-protein interaction	195	4.6	1.1
Physico-chemical process	550	2.7	1.3
BRENDA	2,222	3.3	1.2
Human disease	19,137	5.5	1.0
GO	200,002	4.1	1.4

Table 1: Evaluation of GlycO with respect to. other biomedical ontologies

Pathways can be queried using GlycO, even though they are not explicitly defined the way they are in some databases. A pathway is essentially a sequence of reactions that lead from one chemical compound to another. The advantage of our representation is, that any path between compounds can be shown, by traversing relationships, even if these compounds are not explicitly assigned to a specific pathway, given that all the reactions that are involved are formalized in the ontology. This makes the representation of pathways in the ontology more flexible than that in many databases. Figure 5 shows the GlycO representation of some steps in the N-Glycan biosynthesis pathway.

Another application that requires sophisticated algorithms on databases is described in [10]. The different glycan trees that the authors identify are inherently encoded in the canonical residues and links and can thus easily be queried as well as visualized.

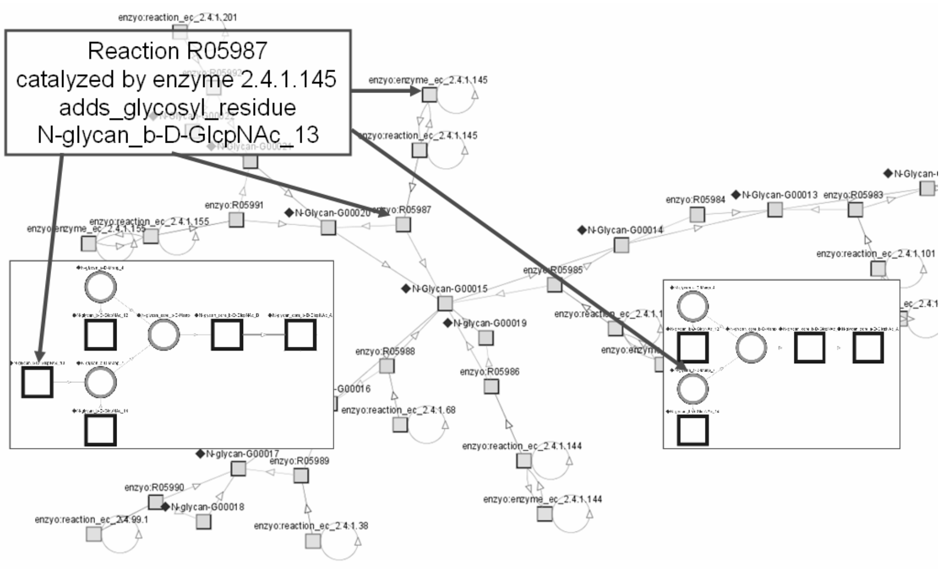

Reaction R05987
catalyzed by enzyme 2.4.1.145
adds_glycosyl_residue
N-glycan_b-D-GlcpNAc_13
enzyo:reaction_ec_2.4.1.201
enzyo:enzyme_ec_2.4.1.145
enzyo:reaction_ec_2.4.1.145
enzyo:R05991
N-Glycan-G00020
enzyo:R05987
enzyo:reaction_ec_2.4.1.155
enzyo:enzyme_ec_2.4.1.155
N-Glycan-G00015
N-glycan-G00019
enzyo:R05988
enzyo:R05986
-G00016
enzyo:reaction_ec_2.4.1.68
N-glycan-G00017
enzyo:R05989
enzyo:R05990
N-glycan-G00018
enzyo:reaction_ec_2.4.99.1
enzyo:reaction_ec_2.4.1.38
enzyo:reaction_ec_2.4.1.144
enzyo:enzyme_ec_2.4.1.144
N-Glycan-G00014
enzyo:R05984
N-Glycan-G00013
enzyo:R05983
enzyo:R05985
enzyo:reaction_ec_2.4.1.101

Figure 5: A part of the N-Glycan biosynthesis pathway as encoded in GlycO. For better visibility, only few relationship types are visualized. N-glycan_b-D-GlcpNAc_13 is the beta-D-GlcpNAc residue number 13 as enumerated in the GlycoTree model.

5. Conclusion

GlycO is not only a vocabulary or a schema meant for database integration, but provides a rich description of the knowledge in the glycoproteomics domain, semantically describing interactions and functions of structures and their substructures as well as their synthesis.

By semantically modeling the structure of molecules with reusable canonical instances, we can evaluate the hypothesis that larger structures exhibit properties and functions that can partially be inferred from the knowledge of the properties and functions of their substructures. The GlycO schema allows a glycan structure to be represented as more than the sum of its parts, paving the way for the identification of the molecular basis for emergent properties. To our knowledge is GlycO the first ontology that models its domain in such detail as described. The formalization of this knowledge allows immediate access to information that so far is only available through specialized tools and algorithms that work on the textual representation in the various biochemistry databases. It was shown that with a sufficiently rich schema alongside trusted sources, automatic extraction, modeling and classification of high-quality instance data is possible.

In the context of this modeling, mereological problems were encountered and addressed. By promoting some of the relationships in the ontology to first class objects, recursive definitions of these relationships allow their expression on different levels of granularity.

6. Acknowledgement

This work is part of the Integrated Technology Resource for Biomedical Glycomics (5 P41 RR18502-02), funded by the National Institutes of Health National Center for Research Resources. Donation by Semagix of its Freedom platform for semantic application development is also acknowledged.

References

[1] Bittner T, Smith B. *A theory of granular partitions*. Foun dations of Geographic Information Science, M Duckham et al. (eds.). London: Taylor & Francis, 2003: 117-151.

[2] Bohne-Lang A.; Lang E.; Forster T.; von der Lieth CW. *LINUCS: linear notation for unique description of carbohydrate sequences*. Carbohydr Res. 336:1-11, 2001.

[3] Doubet, S. and Albersheim, P. *CarbBank*. Glycobiology, 2, 1992

[4] EnzyO. lsdis.cs.uga.edu/projects/glycomics/enzyo/

[5] Gangemi, A.; Catenacci, C.; Ciaramita, M.; Lehmann, J. *A theoretical framework for ontology evaluation and validation*. Proceedings of the 2nd Italian Semantic Web Workshop, Trento, Italy, 2005.

[6] Gene Ontology Consortium. *Gene Ontology: Tool for the Unification of Biology*. Nature Genetics, 25:25-29, 2000.

[7] Genesereth, M. R., and Fikes, R. E. Knowledge Interchange Format, Version 3.0 Reference Manual. Technical Report Logic-92-1, Computer Science Department, Stanford University, 1992

[8] Guarino, N. *Toward a formal evaluation of ontology quality*. IEEE Intelligent Systems, 19(4), 2004

[9] Guarino, N. and Welty, C. *Evaluating Ontological Decisions with OntoClean*, Comm. ACM, 45(2), 2002, pp. 61–65.

[10] Hashimoto, K., Kawano, S., Okuno, Y., and Kanehisa, M. *Global Tree of Known Carbohydrate Structures to Analyze Biosynthetic Pathways*. 15th International Conference on Genome Informatics (GIW2004), December 2004.

[11] Horrocks, I.; Patel-Schneider, P.F. and van Harmelen, F. *From SHIQ and RDF to OWL: the making of a Web Ontology Language*, Journal of Web Semantics 1(1): 7-26 (2003)

[12] Kanehisa, M. and Goto, S. *KEGG: Kyoto Encyclopedia of Genes and Genomes*, Nucleic Acids Research, 2000, Vol. 28(1)

[13] Klyne, G and Carroll, J. Resource Description Framework (RDF): Concepts and abstract syntax. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/ (2004)

[14] Lenat, D. and Guha, R.V. *Building Large Knowledge-Based Systems*: Representation and Inference in the Cyc Project, Addison-Wesley. 1990.

[15] Loß, A.; Bunsmann, P.; Bohne, A.; Loß, A.; Schwarzer, E.; Lang, E. and Von der Lieth, C.-W. *SWEET-DB: an attempt to create annotated data collections for carbohydrates*, Nucleic Acids Research, 2002, Vol 30(1)

[16] Noy, N.F. and McGuinness, D.L. *Ontology Development 101: A Guide to Creating Your First Ontology*. Knowledge Systems Laboratory, 2001

[17] OBO: Open Biomedical Ontologies. http://obo.sourceforge.net

[18] Raymond, J.W. and Willett, P. *Maximum common subgraph isomorphism algorithms for the matching of chemical structures*, Journal of Computer-Aided Molecular Design, 16(7), 2002.

[19] Sahoo, S.S.; Thomas, C.J. Sheth, A.P.; Henson, C; York, W.S. *GLYDE - An expressive XML standard for the representation of glycan structure.* Carbohydrate Research, 340(18), 2005

[20] Sheth, A.; Bertram, C.; Avant, D.; Hammond, B.; Kochut, K.; Warke, Y. *Managing Semantic Content for the Web*, IEEE Internet Computing, July/August 2002.

[21] Sheth, A.P.; Arpinar, I.B. and Kashyap, V. *Relationships at the Heart of Semantic Web: Modeling, Discovering, and Exploiting Complex Semantic Relationships*, in Enhancing the Power of the Internet: Studies in Fuzziness and Soft Computing, M. Nikravesh, L. A. Zadeh, B. Azvine, R.R. Yager (Eds), Springer-Verlag, 63-94, 2004

[22] Smith, B.; Ceusters, Werner; Klagges, B.R.E.; Köhler, J; Kumar, A.; Lomax, J.; Mungall, C.; Neuhaus, F.; Rector, A.L. ; Rosse, C. *Relations in Biomedical Ontologies*. Genome Biology, 6, 2005

[23] Soldatova, L.N. and King, R.D. *Are the current ontologies in biology good ontologies?* Nature Biotechnology 23, 2005

[24] SUMO: http://ontology.teknowledge.com/

[25] Takahashi, N. and Kato, K. *GlycoTree*, Trends in Glycoscience and Glycotechnology, 15, 2003.

[26] Samir Tartir, I. Budak Arpinar, Michael Moore, Amit Sheth, Boanerges Aleman-Meza. *OntoQA: Metric-Based Ontology Quality Analysis*, IEEE ICDM 2005 Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources. Houston, Texas, November 27, 2005